Math. Appl. **8** (2021), 1–13

ASSESSING FAIR POLICING IN AUSTIN, TX

QIUYI WU*, DAVID SKRILL* AND CUONG PHAM

Abstract. This report demonstrates disparities by race in traffic stops by the Austin Police Department. After exploratory analysis, we assess various models and statistics derived from the hit rate and using the Veil of Darkness. We conclude with a Bayesian hierarchical model that produces officer-level posteriors for the hit rate.

1. Introduction

This paper investigates racial disparities in traffic stops by the Austin Police Department. Using data available from Austin Open Data, a Texas government-run data portal, and from the Stanford Open Policing Project, we evaluate these disparities using models derived from the "hit rate" and the effect of the "veil of darkness", two often-cited methods for assessing fair policing. Our main report consists of three parts. First, we conduct an exploratory data analysis to get a big picture of policing in Austin. Second, we use various modeling strategies to assess the severity of racial disparities. Third, we propose a measure of fairness based on the differences in the posterior median hit rate among individual police officers.

2. Available Data

The primary data set is from the Stanford Open Policing Project [1] (from here on referred to as the Stanford data). This data set record stops made by Austin Police Department (APD) over a roughly ten year period (2006.01.01 - 2016.06.30) and contains information such as the date of stops, subject's race, whether the subject was searched or frisked, and whether any contraband were found. Notably, this data lacks information about the time or place of the stops. Because the 2016 data is incomplete, we focus on the data for which we have complete years (2006-2015) for the first two prats of the analysis; this contains 463,944 stops. We note that the distribution of stops per officer has an extremely long tail.

Keywords: Fairness, policing, traffic stops, racial disparities.

The research was part of data analytics competition in 2021 Upstat Annual Conference, and was supported by The American Statistical Association, and RIT Department of Criminal Justice, Fox Pest Control – Rochester, and Wegmans.

* These authors contributed equally to this work.

[1] Stanford Open Policing Project (OPP): https://openpolicing.stanford.edu/data/

We also make use of data from the 2019 Racial Profiling report, available from Austin Open Data (and hereafter referred to as the RP data). This data set contains similar information as the Stanford data, with additional information about the event time, location, and officer race. Notably, the race of the subject is missing from this data.

Lastly, we use US census demographic data. Specifically, we use 2017 5-year American Community Survey zip-code-level data with the Stanford data, and 2019 census population data for 2019 Austin RP data. In addition, we also refer to the racial profiling reports published the Austin Police Department.

3. Exploratory Analysis

We first examine the count of stops by race during 2006-2015 (Table 1), using the Stanford Data. It is notable that over half of the stops involved were of white subjects, about four times the number of stops of Black people. According to 5-year census data, the white population in Austin (445,269) is almost 7 times than the black population (66,724). Examining figure 1, we can see that at least for Black, Hispanic and white drivers, the annual trends are very different by race.

We see that the proportion of stops of white people declined from the beginning of the study period to the end, whereas the proportion of stops of Hispanic and Black people rose and remained roughly constant, respectively.

Driver Race	Counts	Proportion
Asian/Pacific	11658	0.033
Black	52381	0.147
Hispanic	765707	0.215
Other	2105	0.006
Unknown	2622	0.007
White	211588	0.593

Table 1. Proportion of stops by race during 2006-2015.

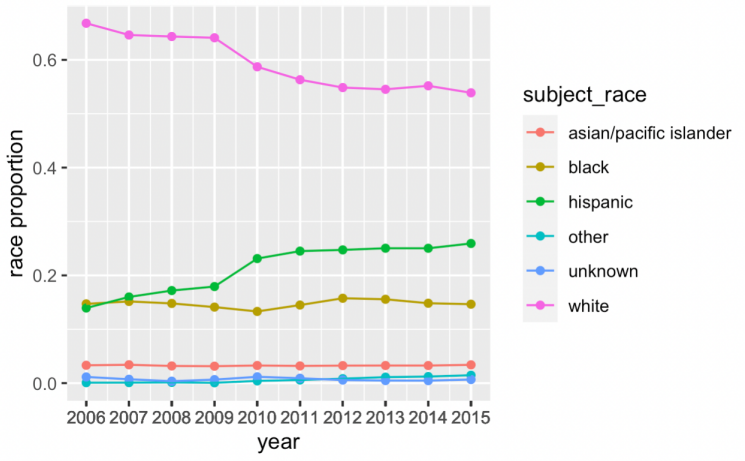

Figure 1. Race proportion in each year.

3.1. Benchmark Test

As shown in Table 1, the number of white drivers stopped is four times the number of black drivers stopped. However,the white population in Austin is over 6 times larger than the Black population. Hence, it is more useful to consider the rate at which different populations are stopped; we define the *stop rate* for a given race as follows.

$$\text{Stop Rate}_i = \frac{\text{Number of Stops for Race } i}{\text{Population of Race } i}$$

From Table 2, we can see black drivers are stopped with a rate much higher than the drivers of other races. We can investigate this further by looking at other benchmarks such as *search rate* and *frisk rate*, defined as follows.

$$\text{Search Rate}_i = \frac{\text{Number of Stopped People Who Were Searched for Race } i}{\text{Number of Stops for Race } i}$$

$$\text{Frisk Rate}_i = \frac{\text{Number of Stopped People Who Were Frisked for Race } i}{\text{Number of Stops for Race } i}$$

As shown in Table 2, Black and Hispanic drivers are searched with a rate 3 times higher than white drivers, and almost 8 times higher than Asian/Pacific drivers. The Black and Hispanic drivers are also frisked at a rate much higher than the drivers of other races.

Although this is strong evidence that stopped Black and Hispanic drivers are searched and/or frisked at a higher rate than stopped white drivers, it is insufficient for making claims about the presence of discriminatory policing practices. It

Driver Race	Counts	Population	Proportion	Stop Rate	Search Rate	Frisk Rate	Hit Rate
asian/pacific	11658	63752	0.033	0.183	0.015	0.011	0.188
black	52381	66724	0.147	0.785	0.092	0.039	0.254
hispanic	765707	316709	0.215	0.242	0.086	0.044	0.323
white	211588	445269	0.593	0.475	0.031	0.021	0.318

Table 2. Stop rates, search rates and frisk rates during 2015.

is necessary to investigate whether different populations are disproportionately stopped relative to their rates of violating the law.

3.2. Outcome Test

In order to assess this, we shall look at the result of searches to see at what rate searched drivers are found to be doing something illegal. In this section, we define a successful search as one that uncovers contraband, and we define the hit rate as the proportion of searches that are successful.

$$\text{Hit Rate}_i = \frac{\text{Number of Times Contraband is Uncovered for Race } i}{\text{Number of Searched People for Race } i}$$

If racial groups have different hit rates, it can be taken as evidence of discriminatory policing: if the hit rate for group i is lower than the hit rate for group j, it may be the case that there is a lower evidenciary standard used when deciding whether to search or frisk a person belong to group i relative to that for a person belonging to group j. As shown in the last column in Table 2, the hit rates for Black and Asian/Pacific drivers are lower than for white and Hispanic drivers, indicating police may have a lower threshold of evidence when searching Black or Asian/Pacific drivers.

3.3. Veil of Darkness Test & Fisher's Exact Test

According to Grogger and Ridgeway, the "Veil of Darkness" test can be used to help assess bias in stop decisions. The hypothesis of this test states that officers who are engaged in racial profiling are less likely to be able to identify a driver's race after dark than during daylight. Under this hypothesis, if stops made after dark had a smaller proportion of a given population than stops made during daylight, it could be evidence of racial profiling.

Traditional uses of the Veil of Darkness rely on knowing the individual races of those stopped. Because neither of our data sets contain both driver races and stop time information, we propose an indirect application at the zip code level. Specifically, using zip-code level demographic data, we label zip codes as either a White Majority Area (WMA, having a population that is more than 50% white) or a Black Majority Area (BMA). For simplicity of the analysis, we consider only zip codes that have either majority white or majority Black populations. We then

consider differences in the number of stops in 2019 between WMAs and BMAs.

In order to accurately distinguish the daytime and nighttime, we computed the daily subset and dusk time for Austin in 2019. In Table 3 we can see that the earliest sunset in 2019 was at around 17:32 in early December; it was fully dark in 26 minutes. The latest sunset time was around 20:38 late June and it was fully dark after 28 minutes.

Date	Sunset	Dusk	Sunset Minute	Dusk Minute
2019-12-02	17:31:42	17:57:48	1051	1077
2019-12-01	17:31:45	17:57:48	1051	1077
2019-06-30	20:37:58	21:05:27	1237	1265
2019-06-29	20:37:56	21:05:27	1237	1265

Table 3. Minimum and maximum dusk time during the 2019 in Austin.

We denote the stops happening before sunset as a daytime stop, and the stop happening after dusk as a nighttime stop. We exclude stops occurring between sunset and dusk.

	Day	Night	Total
BMA	124	126	250
WMA	2937	2216	5153
Total	3061	2342	5403

Table 4. Contingency Table.

We form Table 4 and treat the rows as independent binomial samples. Of $n_1 = 250$ recorded stops in black dominated area, 124 stops happened during the daytime, a proportion of $p_1 = 124/250 = 0.496$. Of $n_2 = 5153$ recorded stops in white dominated area, 2937 stops happened during the daytime, a proportion of $p_2 = 2937/5153 = 0.570$. The sample difference of proportions is 0.074. Using Fisher's exact test to test the null hypothesis of independence of the rows in the following 2×2 contingency table, we obtain a P-value of 0.02, indicating strong evidence against the null hypothesis that the rows are independent, i.e., that the distribution of daytime and nighttime stops are independent of whether a stop occurs in a majority white zip code or a majority Black zip code.

The above test is evidence of different policing practices that depend on the demographic composition of the area being policed. However, we recall that the hypothesis underlying the Veil of Darkness is that policing practices change depending on whether an officer can *see* the race of a driver. While the above is consistent with this hypothesis, it is also consistent with other possibilities, e.g.,

that the APD patrols WMAs less at night. We are hampered by the lack of individual-level race and time data necessary to make a traditional Veil of Darkness approach possible. However, we still believe that this is an interesting and worthwhile result.

3.4. Hit Rate and Causal Issues

In our analysis of observational data, it is important to consider unmeasured confounders. For example, one can argue that it would be unsurprising if more officers patrolled areas with higher crime rates; crime rates are known to be correlated with income and demographic factors. Therefore, if higher-crime neighborhoods have higher minority populations, we would expect to see more minority traffic stops. Although this still exposes problems in Austin, it could be interpreted as a problem of economic segregation, not traffic fairness.

To overcome this problem, we examine the hit rate in greater detail Section 4. We argue that given a person is being searched, the probability of finding contraband items should be equal among all races, regardless of the neighborhood that the search conducted. We want to emphasize that using hit rate does not eliminate all unmeasured confounders, but it helps mitigate the problem.

4. Modeling

4.1. Logistic Regression

Our descriptive analysis shows that black people in Austin seem to be more likely to be stopped by the police. We are interested in the following question: given a person is stopped, what factors may impact the likelihood of that person being frisked? To investigate this, we fit a logistic regression modeling the binary outcome frisk as dependent on race, age, and sex (see Table 5 for summary statistics).

	Nobs	Nmis	Uniq.	Mean	SD	min	25%	50%	75%	max
Subject Age	480091	3164	94	37.98	13.82	10	26	36	48	103
Subject Sex	482881	374	2	0.30	0.46	0	0	0	1	1
Frisk Performed	483255	0	2	0.02	0.15	0	0	0	0	1
Search Conducted	483255	0	2	0.04	0.20	0	0	0	0	1
Person Searched	483255	0	2	0.03	0.18	0	0	0	0	1
Vehicle Searched	483255	0	2	0.02	0.15	0	0	0	0	1
Contraband Found	19256	0	2	0.25	0.42	0	0	0	0	1
Drugs Found	19256	0	2	0.01	0.12	0	0	0	0	1
Weapons Found	19256	0	2	0.05	0.21	0	0	0	0	1
Frisk Performed	19256	0	2	0.51	0.5	0	1	1	1	1

Table 5. Selected Summary Statistics.

$$\text{Logit}[\Pr(\text{Being Frisked})] = \beta_0 + \beta_1 \text{Race} + \beta_2 \text{Age} + \beta_3 \text{Sex}$$

term	estimate	std.error	p.value
(Intercept)	-2.984	0.102	0.000
Black	1.503	0.100	0.000
Hispanic	1.310	0.099	0.000
White	0.719	0.099	0.000
Other	0.617	0.180	0.001
Unknown	0.647	0.183	0.000
Age	-0.046	0.001	0.000
Sex (female)	-1.643	0.036	0.000

Table 6. Logistic model for frisk rate vs. race, age, and sex.

We also investigate how likely it is that contraband is found, given that a search is performed. This is equivalent to calculating hit rate defined previously. We argue that if racial bias does not exist, the hit rate should be equal for all races. In other words, we expect to find that race is not an essential factor in the model.

$$\text{Logit}[\Pr(\text{Contraband found})] = \beta_0 + \beta_1 \text{Race}$$

term	estimate	std.error	p.value
(Intercept)	-1.988	0.222	0.000
Black	0.899	0.225	0.000
Hispanic	0.941	0.224	0.000
White	0.826	0.224	0.000
Other	0.796	0.355	0.025
Unknown	-0.209	0.416	0.616

Table 7. Logistic model for contraband found vs. race.

We also break down the type of contraband found into three categories: Drugs, Weapons, and Others. We also fit a logistic regression model for each of these categories with `Race` as the sole independent variable.

Contraband found	Drugs	Weapons	Others
(Intercept)	-5.24*** (1.00)	-3.61*** (0.45)	-2.38*** (0.26)
Black	1.10(1.01)	0.32(0.46)	0.99***(0.26)
Hispanic	1.21(1.01)	0.36(0.46)	1.04*** (0.26)
White	0.73(1.01)	0.98* (0.46)	0.72** (0.26)
Other	0.97(1.42)	0.47(0.74)	0.87* (0.40)
Unknown	-11.33(280.85)	0.04(0.85)	-0.23(0.53)

Table 8. Contraband Found Break Down (***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).

From Table 6 we see that a Black or Hispanic person is more likely to be frisked than a white person: the estimated odds of a Black person being frisked are 2.22 time those for a white person; this odds ratio a Hispanic person is 1.8. Asian people are the least likely to be frisked. From table 7 and table 8, contraband items are more likely to be found for Hispanic and Black people. White people are more likely to be found with weapons and Black and Hispanic people are more likely to be found with contraband items that are neither drugs or weapons.

4.2. Bayesian Modeling

4.2.1. Investigating the Hit Rate. The "hit rate," defined in this section as the proportion of times an officer finds contraband given that a frisk has been performed, is a widely-used measure for assessing potentially-discriminatory policing. The hit rate can be thought of as a proxy for "evidence" when an officer decides whether to conduct a search or a frisk; a lower hit rate for a particular segment of the population can signal that an officer has a lower threshold of evidence when policing that population segment. In the following analysis, we examine the hit rate at the officer level. Because the analysis requires that officers have stopped and frisked all races under consideration, we restrict the analysis to only White, Black, and Hispanic subject races and to officers with 18 or more frisks, corresponding to roughly the 90th percentile (see Figure 2).

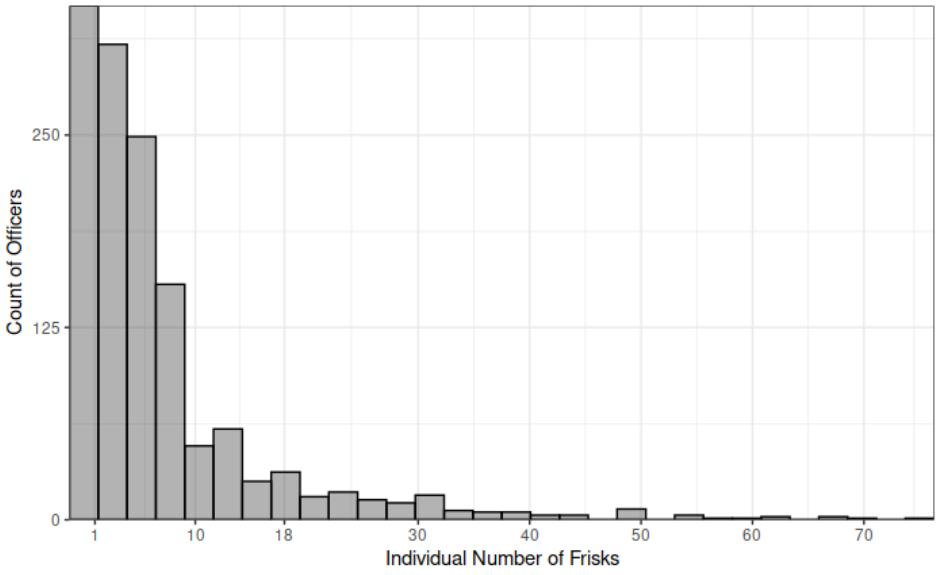

Figure 2. Distribution of Frisks among officers. Note: the very long tail (with maximum exceeding 10,000) is not shown for ease of visualization.

Figure 3 shows the hit rates for individual officers. An officer with an identical hit rate for white and minority subpopulations would be on the 45-degree line. Visually, it is difficult to determine a systematic trend, although it is clear that particular officers have hit rates that differ substantially by subpopulaton. It should be noted that the hit rate is highly variable with small sample sizes.

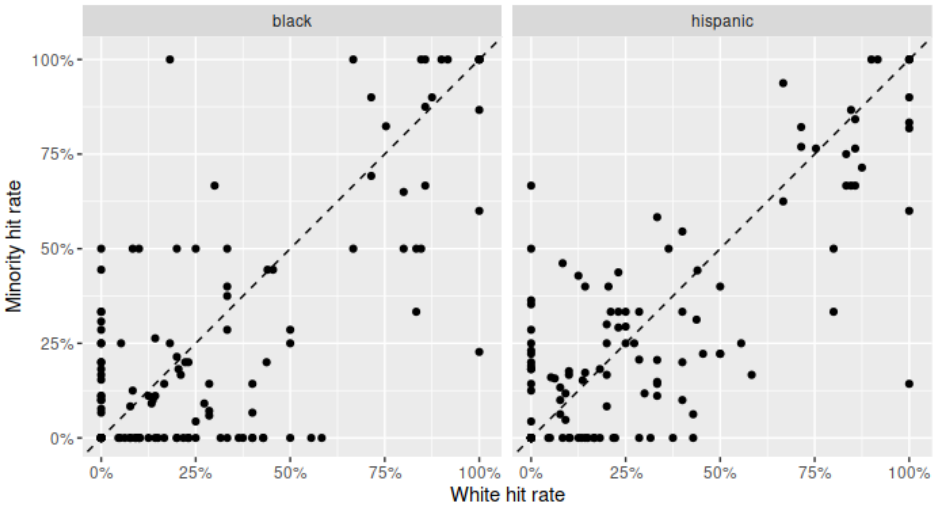

Figure 3. White v. Black and White v. Hispanic Hit Rates.

We proceed using a Bayesian hierarchical model. Under this model, we treat individual officers as belonging to a population and we seek to model both the hit rates of the officers and the variation of this population. This permits partial pooling, by which individual hit rates are biased towards the population average by an amount determined by the estimate of the population. For each officer, we consider three hit rates, one each for white, Black, and Hispanic subpopulations. We accomplish this by fitting separate Bayesian logistic mixed effects models for each race, each with a weakly informative Normal prior on the log-odds with mean -1.3 and standard deviation 1.

Specifically, let θ_{jr} be the hit rate for for officer j and race r, y_{jr} be the number of hits, and K_{jr} the number of frisks. In the following, because we fit separate models, we assume for example $r = black$ and drop the r subscript. Assuming each officer's searches are independent Bernoulli trials

$$p(y_j|\theta_j) = \text{Binomial}(y_j|K_j, \theta_j)$$

We reparametrize the model in terms of the log-odds, α:

$$\alpha_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j}$$

We set a weakly informative prior centered at $\alpha_j = -1.3$, corresponding to $\theta_j \approx 0.2$. The model is therefore

$$p(y_j|K_j, \alpha) = \text{Binomial}(y_j|K_j, \text{logit}^{-1}(\alpha_j))$$

We proceed using `stan_glmer` and the default prior on the covariance matrix. The result includes a posterior for each officer; we may transform from the log-odds back to hit rate to to obtain a posterior for the the hit rate for each officer. We model each race separately, and so obtain three posteriors for each officer. Table 9 shows the posteriors for the first several officers (rows) for each of the three races (columns).

Race	White			Hispanic			Black		
Officer ID	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
01db7098a7	0.003	0.055	0.317	0.021	0.198	0.617	0.065	0.214	0.450
020579eaad	0.021	0.095	0.248	0.002	0.033	0.203	0.047	0.162	0.359
02b0803fe3	0.003	0.056	0.324	0.111	0.242	0.420	0.037	0.174	0.423
0329f48f95	0.128	0.523	0.900	0.057	0.521	0.959	0.665	0.874	0.974
068ff01d47	0.054	0.188	0.415	0.032	0.332	0.833	0.082	0.268	0.556

Table 9. Posterior intervals for several officers for three races.

The effects of partial pooling are evident in Figure 4: the posterior medians are biased towards the population average. Practically, this means that observed hit rates equal to zero have posterior medians that are small but positive, and perfect (or near-perfect) observed hit rates have somewhat smaller posterior medians.

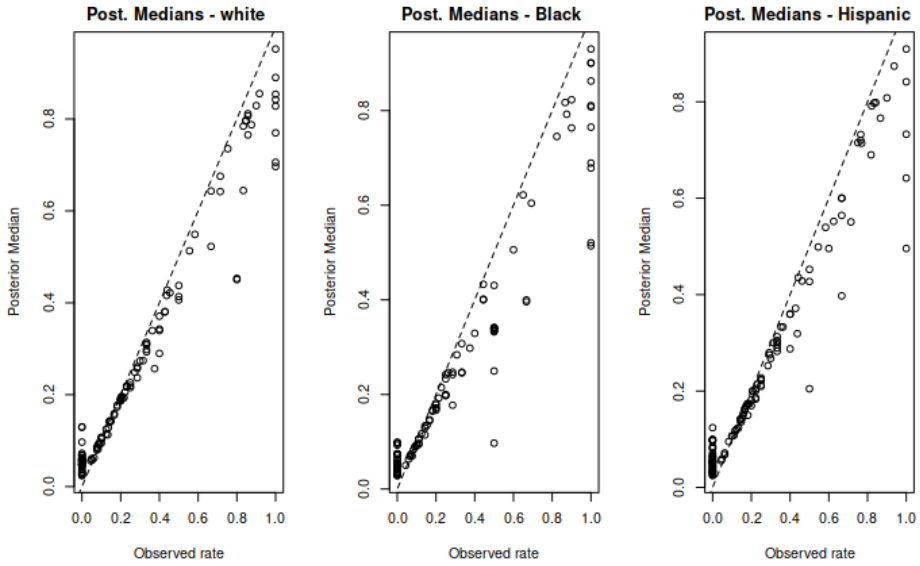

Figure 4. Posterior medians for three races: partial pooling biases the medians towards th population average.

To "operationalize" fairness, we devised a measure by which the above posteriors can be converted into a rough "fairness score." Because an officer that uses the same evidence threshold when deciding whether to frisk a subject regardless of race should have roughly equal hit rates for all three subpopulations, we reason that such an officer should have posterior medians that are close to each other for the three subpopulations. So, one can calculate a simple sum of squares statistic for each officer. Specifically, letting m_{jr} be the posterior median for officer j and race r, the sum of squares statistic S_j is

$$S_j = \sum_r (m_{jr} - \bar{m}_j)^2$$

where $\bar{m}_j$ is the average of the three medians. Of course, this measure disregards all other information that could be gleaned from the posterior; an alternative might calculate the overlap between the posterior densities.

Figure 5 shows the results of this procedure, and Table 10 reports the five highest scores. Examination of Table 10 confirms that partial pooling is working as expected. For example, officer 1504c3bc16 has a perfect observed hit rate when frisking white drivers, but the sample size is small; hence, the posterior median is shrunk significantly. In contrast, the same officer has a much larger sample size for frisking Black drivers, and the posterior median is very close to the observed hit rate. The same observation can be made about officers 3392a495a3, 50f70c6ecb, and dd9c1003d5, who all had observed hit rates of 0 among frisked Black drivers.

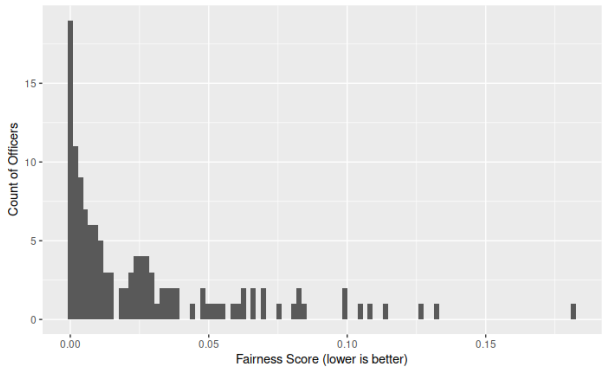

Figure 5. Fairness scores calculated from the posterior medians. An officer with identical posterior medians would have a score of 0.

		Observed			Posterior			Counts		
Officer ID	Score	White	Black	Hispanic	White	Black	Hispanic	White	Black	Hispanic
1504c3bc16	0.1818	1.000	0.227	0.143	0.697	0.215	0.142	2	22	7
50f70c6ecb	0.1314	0.583	0.000	0.167	0.549	0.064	0.162	12	3	18
bab7c2acaf	0.1274	0.833	0.333	0.750	0.644	0.247	0.715	7	3	20
dd9c1003d5	0.1137	0.556	0.000	0.250	0.513	0.043	0.210	9	6	4
3392a495a3	0.1087	0.400	0.000	0.545	0.371	0.046	0.499	10	5	11

Table 10. Highest 5 scores.

5. Conclusion

In this study, we evaluate the fairness of traffic stops during the past two decades through three tests, namely benchmark test, outcome test and veil of darkness test. We found that the racial disparity in policing exists and is present in different scales spatially and temporarily. We also explore the causal confounding issues through logistic regression, finding that black and Hispanic people are more likely to be frisked and found with contraband items that are neither drugs or weapons.

Through the investigation of the hit rate via Bayesian hierarchical modeling, we obtained posteriors for the hit rate for each officer in a subset of the data. Using the medians for these posteriors, we devised a "fairness score," a tool we believe could be used to identify officers with racially disparate patterns of traffic stops.

References

[1] Grogger, Jeffrey, and Greg Ridgeway. "Testing for racial profiling in traffic stops from behind a veil of darkness." Journal of the American Statistical Association 101.475 (2006): 878-887.
[2] Carpenter, Bob, J. Gabry, and B. Goodrich. "Hierarchical partial pooling for repeated binary trials." (2016).

[3] Pierson, Emma, et al. "A large-scale analysis of racial disparities in police stops across the United States." Nature human behaviour 4.7 (2020): 736-745.
[4] Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. "The problem of infra-marginality in outcome tests for discrimination." The Annals of Applied Statistics 11.3 (2017): 1193-1216.

Qiuyi Wu*, Department of Biostatistics and Computational Biology, University of Rochester, *e-mail*: Qiuyi_Wu@URMC.Rochester.edu

David Skrill*, Department of Biostatistics and Computational Biology, University of Rochester, *e-mail*: David_Skrill@URMC.Rochester.edu

Cuong Pham, Department of Biostatistics and Computational Biology, University of Rochester, *e-mail*: Cuong_Pham@URMC.Rochester.edu